Outlier Detection Techniques for Data Mining

Fabrizio Angiulli
University of Calabria, Italy

INTRODUCTION

Data mining techniques can be grouped in four main categories: clustering, classification, dependency detection, and outlier detection. Clustering is the process of partitioning a set of objects into homogeneous groups, or clusters. Classification is the task of assigning objects to one of several predefined categories. Dependency detection searches for pairs of attribute sets which exhibit some degree of correlation in the data set at hand.

The outlier detection task can be defined as follows: "Given a set of data points or objects, find the objects that are considerably dissimilar, exceptional or inconsistent with respect to the remaining data". These exceptional objects as also referred to as outliers.

Most of the early methods for outlier identification have been developed in the field of statistics (Hawkins, 1980; Barnett & Lewis, 1994). Hawkins' definition of outlier clarifies the approach: "An outlier is an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism". Indeed, statistical techniques assume that the given data set has a distribution model. Outliers are those points that satisfy a discordancy test, that is, that are significantly far from what would be their expected position given the hypothesized distribution.

Many clustering, classification and dependency detection methods produce outliers as a by-product of their main task. For example, in classification, mislabeled objects are considered outliers and thus they are removed from the training set to improve the accuracy of the resulting classifier, while in clustering, objects that do not strongly belong to any cluster are considered outliers. Nevertheless, it must be said that searching for outliers through techniques specifically designed for tasks different from outlier detection could not be advantageous. As an example, clusters can be distorted by outliers and, thus, the quality of the outliers returned is affected by their presence. Moreover, other than returning a solution of higher quality, outlier detection algorithms can be vastly more efficient than non ad-hoc algorithms.

While in many contexts outliers are considered as noise that must be eliminated, as pointed out elsewhere, "one person's noise could be another person's signal", and thus outliers themselves can be of great interest. Outlier mining is used in telecom or credit card frauds to detect the atypical usage of telecom services or credit cards, in intrusion detection for detecting unauthorized accesses, in medical analysis to test abnormal reactions to new medical therapies, in marketing and customer segmentations to identify customers spending much more or much less than average customer, in surveillance systems, in data cleaning, and in many other fields.

BACKGROUND

Approaches to outlier detection can be classified in supervised, semi-supervised, and unsupervised.

Supervised methods exploit the availability of a labeled data set, containing observations already labeled as normal and abnormal, in order to build a model of the normal class. Since usually normal observations are the great majority, these data sets are unbalanced and specific classification techniques must be designed to deal with the presence of rare classes (Chawla et al., 2004).

Semi-supervised methods assume that only normal examples are given. The goal is to find a description of the data, that is a rule partitioning the object space into an accepting region, containing the normal objects, and a rejecting region, containing all the other objects. These methods are also called one-class classifiers or domain description techniques, and they are related to novelty detection since the domain description is used to identify objects significantly deviating form the training examples.

Unsupervised methods search for outliers in an unlabelled data set by assigning to each object a score which reflects its degree of abnormality. Scores are usually computed by comparing each object with objects belonging to its neighborhood.

Data mining researchers have largely focused on unsupervised approaches. Most of the unsupervised approaches proposed in the data mining literature can be classified as deviation-based (Arning et al., 1996), distance-based (Knorr & Ng, 1998), density-based (Breunig et al., 2000), and MDEF-based (Papadimitriou et al., 2003).

Deviation-based techniques (Arning et al., 1996) identify as exceptional the subset I_x of the overall data set I whose removal maximizes the similarity among the objects in $I–I_x$.

Distance-based outlier detection has been introduced by (Knorr & Ng, 1998) to overcome the limitations of statistical methods: an object O is an outlier in a data set with respect to parameters k and R if at least k objects in the data set lie within distance R from O. This definition generalizes the definition of outlier in statistics. Moreover, it is suitable in situations when the data set does not fit any standard distribution.

(Ramaswamy et al., 2000), in order to provide a ranking of the outliers, modified the previous definition as follows: given two integers k and n, an object O is said to be the n-th top outlier if exactly $n–1$ objects have higher value for D^k than O, where D^k denotes the distance of the k-th nearest neighbor of the object.

Subsequently, (Angiulli & Pizzuti, 2002) with the aim of taking into account the whole neighborhood of the objects, proposed to rank them on the basis of the average distance from their k nearest neighbors, also called weight.

Density-based methods, introduced in (Breunig et al., 2000), are based on the notion of local outlier. Informally, the Local Outlier Factor (LOF) measures the degree of an object to be an outlier by comparing the density in its neighborhood with the average density in the neighborhood of its neighbors. The density of an object is related to the distance to its k-th nearest neighbor. Density-based methods are useful when the data set is composed of subpopulations with markedly different characteristics.

The multi-granularity deviation factor (MDEF), introduced in (Papadimitriou et al., 2003), is in principle similar to the LOF score, but the neighborhood of an object consists of the objects within an user-provided radius and the density of an object is defined on the basis of the number of objects lying in its neighborhood.

In order to discover outliers in spatial data sets, specific definitions are needed: spatial outliers (Shekhar et al., 2003) are spatially referenced object whose non-spatial attribute values are significantly different from the values of their spatial neighborhood, e.g. a new house in an old neighbourhood.

MAIN FOCUS

We overview several recent data mining techniques for outlier detection. We will focus on distance and density-based methods for large data sets, on subspace outlier mining approaches, and on algorithms for data streams.

Algorithms for Distance- and Density-Based Outliers

Distance-based outlier scores are monotonic non-increasing with respect to the portion of the data set already explored. This property allows to design effective pruning rules and very efficient algorithms.

The first two algorithms for mining distance-based outliers in large data sets were presented in (Knorr et al., 2000). The first one is a block nested loop that runs in time quadratic in the size of the data set. The second one is a cell-based algorithm whose temporal cost is exponential in the dimensionality of the data. These methods do not scale well for both large and high-dimensional data. Thus, efforts for developing scalable algorithms have been subsequently made. (Ramaswamy et al., 2000) present two algorithms to detect the top outliers. The first assumes that the dataset is stored in a spatial index. Since the first method is computationally expensive, they also introduce a clustering-based based algorithm, which has been tested up to ten dimensions.

The distance-based outlier detection algorithm ORCA (Bay & Schwabacher, 2003) enhances the naïve block nested loop algorithm with a simple pruning rule and randomization and exhibits good scaling behavior on large and high dimensional data. ORCA manages disk resident data sets and employs a memory buffer of fixed size. Its I/O cost, that is the total number of disk blocks accesses, my become quadratic due to the nested loop strategy. (Ghoting et al., 2006) use a divisive hierarchical clustering algorithm to partition data in a suitable data structure and then employ the strategy of ORCA on this structure, obtaining improved performances. Differently than ORCA, the whole data set must be accommodated in a main memory resident